

# Web Services-Based Access to Local Clinical Trial Databases: A Standards Initiative of the Association of American Cancer Institutes

Douglas C. Stahl, PhD<sup>a</sup>; Richard M. Evans Jr, BA<sup>b</sup>; Lawrence B. Afrin, MD<sup>c</sup>;  
Richard M. DeTeresa, BS<sup>d</sup>; Dave Ko, BS<sup>a</sup>; and Kevin Mitchell, MS<sup>e</sup>

<sup>a</sup>City of Hope Comprehensive Cancer Center, Duarte, CA

<sup>b</sup>USC/Norris Comprehensive Cancer Center, Los Angeles, CA

<sup>c</sup>Hollings Cancer Center, Medical University of South Carolina, Charleston, SC

<sup>d</sup>Moore's UCSD Cancer Center, La Jolla, CA

<sup>e</sup>University of Pittsburgh Cancer Institute, Pittsburgh, PA

## Abstract

Electronic discovery of the clinical trials being performed at a specific research center is a challenging task, which presently requires manual review of the center's locally maintained databases or web pages of protocol listings. Near real-time automated discovery of available trials would increase the efficiency and effectiveness of clinical trial searching, and would facilitate the development of new services for information providers and consumers. Automated discovery efforts to date have been hindered by issues such as disparate database schemas, vocabularies, and insufficient standards for easy intersystem exchange of high-level data, but adequate infrastructure now exists that make possible the development of applications for near real-time automated discovery of trials. This paper describes the current state (design and implementation) of the Web Services Specification for Publication and Discovery of Clinical Trials as developed by the Technology Task Force of the Association of American Cancer Institutes. The paper then briefly discusses a prototype web service-based application that implements the specification. Directions for evolution of this specification are also discussed.

## Background

Most clinical trial centers today use, to varying degrees, electronic methods of information management. Some centers develop their own databases, while others purchase solutions from commercial vendors (e.g., PhaseForward's ClinTrial<sup>1</sup>). Database schemas range from simple to complex, but at a minimum, most centers' databases contain a common set of basic data elements describing each available trial. Standard data elements generally include the center's identification

number for the trial, the title of the trial's protocol, the name of the principal investigator, the name of the sponsor, and the accrual status.

Although these databases are almost always constructed primarily for centers' internal trial management purposes, there can be no doubt that there exists a substantial external demand for the information contained therein. Patients need to know if a center has a trial for a given condition. Investigators at other centers need to know about potential collaborators and competitors. Sponsors of new trials need to be able to find potential sites with non-competing trials. Regulators and prospective employees need to understand the full extent of a center's trial activity. Even absent the external demand, the centers themselves have decided interests - e.g., increasing patient accrual - in exposing selected internal information.

Meeting the growing demand for accurate and timely clinical trial information is a challenging task for trial centers, and electronic discovery of the trials being performed at an arbitrary center is likewise a challenging task for consumers of this information<sup>2</sup>. Centers have undertaken a wide variety of methods over the years, including advertising via local traditional media channels, production and distribution of brochures, manual submission of trial information to central registries, and, more recently, publication of web pages listing available trials. (In fact, the latter two methods have been combined, as central registries (e.g., CenterWatch<sup>3</sup>, PDQ<sup>4</sup>, ClinicalTrials.gov<sup>5</sup>) have been publishing web pages for several years.) In some centers the trial listings on such web pages may be statically hand-crafted and require manual updating on a periodic basis, but more likely they are dynamically derived from the centers' databases on demand.

Although research centers already employ a number of methods for publicizing their trial information, the difficulties still present today in finding basic trial information speak to the inadequacies of current publicity methods. As a result, unskilled searchers (e.g., some patients) still rely on intermediaries (e.g., their physicians), and skilled searchers at their best are typically limited to accessing web pages of trial listings from central registries and/or individual research centers.

Each of these types of listings has unique handicaps. A central registry may be able to list a large number of trials in a helpful, consolidated, common format, but because of the manual process (electronic or not) commonly used by research centers and sponsors to submit trial information to central registries, central registry listings often out of date. For example, Manheimer et al<sup>6</sup> found that as many as 50% of current Phase III trials for a common disease such as colon cancer were not included in the ClinicalTrials.gov registry. Also, unless significant network infrastructure precautions are taken, central registries can be central points of failure.

There are different handicaps associated with searches of individual center-specific databases. Although the information may be more current, the scope of the listing typically is limited to only the trials being conducted at that center. Thus, the searcher is challenged to find all the centers likely to be of relevance in the search, and must manually interpret the different formats in which each center displays its clinical trial information.

It should also be noted that current global search engines (e.g., Google<sup>7</sup>) do not bridge the gap between central registries' currency deficiencies and trial center's scope deficiencies because not only do search engines have a lag time<sup>8</sup> (typically weeks to months) in re-indexing any given site, but they also exclude from their indices the types of dynamically generated pages<sup>9</sup> which many centers use to produce their trial listings.

An optimal trial search method would examine all center-specific clinical trials databases in parallel, maximizing both the currency and the scope of the search and avoiding central points of failure.

Practically speaking, initial implementation of such a method requires the global adoption of at least two sets of standards: (1) a standard format for reporting a center's trial information, and (2) a standard method for distributing queries and reports. The initial web service specification described in this paper is

designed to incorporate these standards as they are developed.

## History of the Initiative

The Association of American Cancer Institutes (AACI)<sup>10</sup> was established in 1959 to promote common interests of the nation's leading academic and freestanding cancer centers. Approximately half of the 80 participating centers are NCI-designated, and all are dedicated to cancer research, treatment, prevention, early diagnosis, and education. In January 2001, members from 15 AACI institutions were organized as a Technology Task Force (Table 1) to develop recommendations and describe best practices in cancer informatics and corresponding technology infrastructure requirements. After several months of collaboration, the group's recommendations were presented to the AACI and the NCI in October 2001.

A key recommendation involved the development of integrated data models and information systems for clinical oncology research. Toward this end, Drs. Joyce Niland (City of Hope Cancer Center) and Randolph Miller (Vanderbilt-Ingram Cancer Center) developed a call for participation and a survey to identify AACI centers with well-developed cancer research information models and a willingness to collaborate. In January 2002, task force members reviewed 25 completed surveys and selected 17 centers to participate in the next phase of activities. Representatives from 12 of the centers, along with Dr. Kenneth Buetow from the NCI's Center for Bioinformatics, met at City of Hope in April 2002 to exchange ideas and develop plans for continued development. A complete summary of this meeting and all related documents is available at <http://www.aaci-cancer.org/informatics>.

Participants at the two-day meeting agreed to design and deploy an Internet-based prototype application to integrate publicly available protocol summary information across member centers. To meet this objective, the participants divided into two smaller task forces. The Terminology Task Force was charged with the review of existing terminology standards, rectification of terminology inconsistencies across participating centers, and metadata management. The authors, along with members of several other centers, formed a Technology Task Force charged with technology assessment, interface design, and prototype development.

**Table 1.** Task force participants

<b>Institution</b>	<b>April 2002 Meeting Participant</b>	<b>AACI Technology Task Force</b>	<b>AACI Terminology Task Force</b>	<b>Web Service Deployed (as of 3/ 2003)</b>
City of Hope Comprehensive Cancer Center	√	√	√	√
Dana Farber Cancer Institute / Harvard University	√	√	√	
Herbert Irving Comprehensive Cancer Center Columbia University	√		√	
Hollings Cancer Center Medical University of South Carolina		√		√
Lombardi Cancer Center / Georgetown University		√		
Memorial Sloan-Kettering Cancer Center	√		√	
Moore's Cancer Center University of California San Diego	√	√		√
Norris Comprehensive Cancer Center University of Southern California	√	√	√	√
Norris Cotton Cancer Center / Dartmouth	√	√		
Robert H. Lurie Comprehensive Cancer Center Northwestern University	√			
University of Pittsburgh Cancer Institute	√	√		√
University of Wisconsin Comprehensive Cancer Center	√			
Vanderbilt-Ingram Cancer Center	√			
Yale Cancer Center	√		√	

Since June 2002 the Technology Task Force has been collaborating via e-mail and monthly conference calls. This collaboration has produced an initial set of specifications and guidelines (Figure 1) for a platform-independent web service-based method for publishing and discovering certain basic data elements concerning a center's trials. To date, five participating centers have deployed AACI-compliant web services and are able to share data from the centers' clinical trial databases. The locally implemented web services were developed in a variety of languages (e.g. Visual Basic DotNet, Java) and development environments (e.g. Visual Studio, ColdFusion). We have also developed a prototype application which consumes these services and can produce a dynamically generated, consolidated list of these centers' trials. This application is publicly available at <http://opidev.upmc.edu/aaci/project.jsp>.

Current task force efforts include (1) continued enhancement and testing of currently deployed services, (2) working with new centers to adapt and

deploy services, (3) expansion of the service model to include additional clinical trial data elements.

### Future Directions

Recent Internet technology developments, including HyperText Transport Protocol<sup>11</sup> and Extensible Markup Language (XML)<sup>12</sup>, now provide a standard transport method. Another recent, XML-based Internet technology development, Simple Object Access Protocol (SOAP)<sup>13</sup>, permits definition of a standard, machine-interpretable format for reporting a center's trial information.

While these standards and technologies suffice for the initial implementation, as described herein, of a standardized web services-based method for searching individual centers' trial listings, it should be noted that two additional sets of standards would significantly increase the utility of this method.

**Figure 1.** AACI Web Services Specifications & Guidelines, Rev 1.0 11/2002

<u>Current AACI requirements</u>	
Web service accepts string input for keyword searches (not case sensitive).	
Service must accept multiple keyword searches.	
Web service returns XML data of all protocols matching the keyword(s) input.	
Empty (null) search string returns all open protocols.	
<u>Naming Convention</u>	
Service name:	protocols
Method:	getProtocols()
	Parameter: SearchString, maximum 128 characters
Dataset name (root element):	ProtocolSet
Record name (child element):	Protocol
<u>Dataset Characteristics</u>	
Records returned by the web service needs to contain the following fields:	
• prot_num	institution's unique identifier for a given protocol
• prot_title	name of the protocol
• pi	principal investigator for the protocol
• institution	name of the institution.
• institution_zipcode	zip code of the institution
• institution_county	county of the institution
• prot_accrual_status	protocol accrual status
• prot_url	URL for more information on a given protocol

First, standard vocabularies for the data elements in centers' trial listings would greatly aid the machine-interpretability of this information. Trials-related vocabulary standards work is being pursued by a number of investigators and organizations at present, including Health Level Seven (HL7)<sup>14</sup>, Clinical Data Interchange Standards Consortium (CDISC)<sup>15</sup>, and the National Cancer Institute<sup>16</sup>. The NCI has created, and made available for public use, a cancer data standards repository (caDSR). The caDSR is built according to the ISO 11179<sup>17</sup> standard for metadata repositories. The caDSR defines common data elements (CDE's)<sup>18</sup> containing metadata that describe data relevant to cancer clinical trials. We plan to make use of this repository as a data dictionary for our project. We will either match each XML element in our specification with an existing CDE, or develop CDE's in our context in the caDSR if no such match exists.

Second, a standard method (perhaps based on the relatively new, XML-based Universal Description Discovery & Integration (UDDI) protocol<sup>19</sup>) for locating centers which adhere to the first two sets of standards described above would be the final component for supporting massively parallel searches of individual centers' listings. Alternatively, propagation of trial information among local centers'

trial databases (in a manner similar to propagation of news messages amongst Usenet servers via the Network News Transfer Protocol<sup>20-22</sup>) also could be pursued. Either way, the resulting increased scope of trial searches would be of use not only to individual users but also to central registries desirous of improving their currency.

One primary emphasis will be to increase participation in the current project. To date, five centers have deployed an AACI-compliant web service. We hope to increase that number by reducing the effort required to implement a local web service.

Our work has demonstrated that the only site-specific task involved in creating such a web service is the mapping of the XML elements expected by querying applications to actual tables and fields in the site's existing database or databases. In theory then, we should be able to develop a methodology that would limit the upfront work required to participate in this project to the editing of a mapping file. The local mapping file would implement the XML standard for data schemas (XML Schema<sup>23</sup>), and provide the mechanism for associating each XML element with a corresponding CDE in the caDSR as well as mapping each element to a field in a local database.

Furthermore, we want to determine if the development of a distributable version of an AACI-compliant web service that uses such a mapping file to provide the specifics for accessing local data stores will increase project participation. Once the mapping file has been appropriately edited by the local center, it should contain all the information required by the distributed (downloaded) AACI-compliant web service to query the local database(s), and return a fully qualified XML response to any querying application.

As previously mentioned, there are two additional research foci that we will pursue in the short term: (1) establishment of methods (perhaps to be based on the Universal Discovery & Data Integration protocol) for automated discovery of instances of the AACI-compliant web service, and (2) enhancement of our web service specification to include methods for defining publish/subscribe relationships between trial databases to allow for the development of a massively redundant network of local databases which each contain the totality of (exposed) global trial information. These enhancements may be all that is required to rectify the significant shortcomings of even the most comprehensive trial registries currently available.

In conclusion, we have developed a specification, which we propose as a draft standard for the clinical research community, for a basic clinical trial publishing and discovery web service. Five cancer centers, operating in a variety of computing environments, have implemented prototypes of this service. As further proof of concept, we also have implemented an application which consumes these services and thus can produce a dynamically generated, consolidated list of these centers' trials. Further work will (1) improve the ease of service implementation, (2) propose and prototype a method for automated discovery of these services, (3) increase the number of publicly discoverable, standards-based trial data elements, and (4) propose and prototype an architecture for replication of trial information across multiple centers' database so as to further improve trial discoverability.

## References

1. ClinTrial:  
<http://www.phaseforward.com/products/clintrial.htm>
2. Cassileth B. Clinical Trials: Time for action. *J Clin Onc* 2003 Mar 1; 21(5); 765-766.
3. CenterWatch: <http://www.centerwatch.com>

4. PDQ:  
<http://www.nci.nih.gov/cancerinfo/pdq/>
5. ClinicalTrials.gov: <http://clinicaltrials.gov>
6. Manheimer E, Anderson D. Survey of public information about ongoing clinical trials funded by industry evaluation of completeness and accessibility. *BMJ* 2002 Sep 7; 325. Available from <http://bmj.com/cgi/content/abstract/325/7363/528>
7. Google: <http://www.google.com>
8. Search engine reindex lag time:  
<http://www.devwebpro.com/2002/0927.html>  
<http://www.devwebpro.com/2002/0927.html>
9. Search engine dynamic page indexing constraint: <http://www.edu.com/databases.htm>,  
[http://www.netmechanic.com/news/vol4/promo\\_no3.htm](http://www.netmechanic.com/news/vol4/promo_no3.htm),  
<http://www.phpbuilder.com/columns/tim19990117.php3>
10. AACI Website: <http://www.aaci-cancer.org>
11. HTTP: <http://www.w3.org/Protocols/>
12. XML: <http://www.w3.org/XML/>
13. SOAP: <http://www.w3.org/2000/xp/Group/>
14. HL7: <http://www.hl7.org/>
15. CDISC: <http://www.cdisc.org/>
16. NCI: <http://www.nci.nlm.gov/>
17. caDSR/ISO 11179:  
<http://ncicb.nci.nih.gov/core/caDSR/ISO11179>
18. CDE detail:  
<http://ncicb.nci.nih.gov/core/caDSR/ISO11179>
19. UDDI: <http://uddi.org/>
20. <http://www.faqs.org/faqs/usenet/software/part1/>
21. <http://www.faqs.org/rfcs/rfc977.html>
22. <http://www.faqs.org/rfcs/rfc2980.html>
23. XML schemas:  
<http://www.w3.org/2001/XMLSchema>